INVESTIGATION OF THE ROBUSTNESS OF THE STUDENT'S t-TEST UNDER THE VIOLATION OF THE ASSUMPTION OF EQUALITY OF VARIANCES

by

Harry Alben Hadd

United States
Naval Postgraduate School

THESIS

INVESTIGATION OF THE ROBUSTNESS OF THE
STUDENT'S t-TEST UNDER THE VIOLATION
OF THE ASSUMPTION
OF EQUALITY OF VARIANCES

by

Harry Alben Hadd, Jr.

December 1970

Investigation of the Robustness of the Student's t-Test Under the Violation of the Assumption of Equality of Variances

by

Harry Alben Hadd, Jr.
Captain, United States Marine Corps
B.S., United States Naval Academy, 1965

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 1970

ABSTRACT

The robustness of the Student's t-test is investigated under the violation of the assumption of equality of variances. With the aid of computer simulation, Type I and Type II error rates and the resulting statistical inference are studied and the effects of unequal variances on rejection rates and the power of the test are determined. Limits are determined on the degree of violation of the equality of variances that still leads to a satisfactory result when Student's distribution is used.

TABLE OF CONTENTS

I. INTRODUCTION

In investigating the robustness of the Student's t-test, it is necessary to initially discuss the underlying distribution used by the test, the t distribution. Prior to 1908 statistical analysis was greatly dependent on knowing the population variance σ^2 for most procedures. The random variable

$$z = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} \qquad 1\text{-}1$$

was used extensively. To develop z, the hypothesized population mean μ is subtracted from the sample mean $\bar{x}$ and the resulting value is multiplied by the square root of the sample size n and divided by the population standard deviation σ. The statistic z has a normal distribution with mean zero and standard deviation equal to one, N(0,1), if x is distributed normally with mean equal to μ and standard deviation equal to σ^2, i.e., $N(\mu, \sigma^2)$. When x has any distribution other than $N(\mu, \sigma^2)$, then z approaches a N(0,1) as $n \to \infty$ according to the central limit theorem.

In 1908 Gosset, publishing under the pseudonym of "Student", developed a procedure which modified z for instances where the population variance σ^2 was unknown. He estimated σ^2 using the unbiased estimator

$$s_{\bar{x}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad 1\text{-}2$$

Gosset then considered the random variable

$$t = \frac{(x - \mu)\sqrt{n}}{s_x} \qquad 1\text{-}3$$

As Meyer (17) notes, the probability distribution of the random variable t is more complicated than that of z because both the numerator and denominator of t are random variables whereas z is simply a linear function of the random sample $X_1,\ldots,X_n$.

In an effort to obtain the probability distribution of t, Gosset considered these facts:

1. z has a N(0,1) distribution.

2. $v = \sum_{i=1}^{n} (x_1 - \bar{x})^2/\sigma^2$ has a Chi-square distribution with (n-1) degrees of freedom.

3. z and v are independent random variables.

He defined

$$t = \frac{z}{\sqrt{v/d}} \qquad d = n-1 \qquad 1\text{-}4$$

and found the probability density function (pdf) of t as given by

$$h_d(t) = \frac{\Gamma[(d+1)/2]}{\Gamma(d/2)\sqrt{\pi d}} \left[1 + \frac{t^2}{d}\right]^{-(d+1)/2} \quad -\infty < t < \infty \qquad 1\text{-}5$$

where Γ denotes a Gamma function where $\Gamma(n+1) = n! = \int_0^\infty e^{-x} x^n dx$. This distribution is known as the Student's t-distribution with d degrees of freedom.

The pdf h_d is symmetric with a mean of zero and resembles the normal distribution. Dixon and Massey (3) show that even though on the average s_x^2 is equal to σ^2, more than half the time s_x^2 is actually less than σ^2 because of the kurtosis of the distribution of s_x^2. Lindley (14) has proven through a rigorous mathematical argument that as the sample size n becomes large the density of the t distribution tends to have a distribution N(0,1).

Because of its importance, especially as the underlying distribution for the Student's t-test, the t-distribution has been tabulated.

In the problem of testing the hypothesis that the means of two normal populations are equal the most commonly used test is the Student's t-test. The test as developed by Gosset formulates the following random variable:

$$t = \frac{\bar{x} - \bar{y}}{\left[\frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}\right]^{\frac{1}{2}} \left[\frac{1}{n_x} + \frac{1}{n_y}\right]^{\frac{1}{2}}} \qquad 1\text{-}6$$

where n_x, n_y are defined as the sample sizes drawn respectively from normal populations X and Y.

The variables $\bar{x}$ and $\bar{y}$ are the sample means of the populations X and Y respectively and s_x^2 and s_y^2 are the unbiased sample variances of the X and Y populations respectively.

The underlying distribution for this statistic has the same t-distribution as the statistic shown in (1-3) because $\bar{x} - \bar{y}$ is a normal random variable and the entire denominator is a pooling of the sums of the squared deviations from the means of both samples which provides the best unbiased estimate of the common population variance.

To test the hypothesis the absolute value of the t statistic compiled from the samples is compared to a particular value from the t distribution which has associated with it a probability of a more extreme value. Where the observed absolute value of t, $|t_o|$, is greater than the tabulated $|t|$ value, a hypothesis that the two population means are equal, is rejected. However, if the value of the observed $|t_o|$ statistic is less than the tabulated $|t|$ value, the hypothesis is accepted.

In order to use this particular test for equality of means, as intended, the theory requires certain assumptions be met. The first assumption dictates that the random samples drawn from each population must be independent. Secondly, Gosset stated that the underlying populations from which the samples are taken must be normally distributed. The third and seemingly most severe assumption, is that the variances of both populations must be equal.

This paper is concerned with a detailed empirical study of the ability of the t-test to give correct results to the question of whether or not the means of two normal populations are equal when the third assumption of equal variances is violated. The robustness of the t-test, or its ability to withstand this violation of assumption is investigated for various degrees of violation of the assumption of equal variances. Under this condition, certain error rates are investigated. One type of error rate is the fraction of instances the test implies that the means of two normal populations are not equal when in fact they are equal. The second type of error rate is the fraction of instances that the test implies that the means of the two normal populations are equal when in fact they are not equal. The power of the t-test or its ability to detect the difference between two population means, is a function of the second type of error rate and is equal to one minus the fraction of errors of the second kind.

The investigation of these error rates is conducted for both equal and unequal sample sizes and the ratio of the population variances is allowed to vary over a wide range of values.

II. BACKGROUND

A. STATISTICAL INFERENCE AND HYPOTHESIS TESTING

The evaluation of the robustness and power of a test requires some elementary knowledge in the area of statistical inference and especially hypothesis testing. Generally the observations or random samples drawn from one or more populations are arithmetically manipulated by a particular method to obtain information about the underlying populations. This single number calculated from sample data is referred to as a statistic. From this statistic certain inferences can be made about either a particular parameter of a single population under study or whether equality exists between the same parameters of two or more populations.

The t-test falls into the second major area of statistical inference called hypothesis testing. The test is applied to the common statistical problem of determining whether or not the means of two normally distributed populations are equal. The test begins with the hypothesis that the means are equal and then from the value of the statistic, the decision is made whether the hypothesis is accepted or rejected. From the t statistic developed in 1-6 it should be observed that in testing the hypothesis the direct concern is not with determining the actual value of the means of the two distributions but instead in determining whether a difference exists between the two means.

There are certain basic properties that any method used for hypothesis testing must be required to possess. The first property is that when any hypothesis test is used there should exist only a small probability that the results obtained from the method lead to an erroneous conclusion. In other words, in the case of the t-test, if indeed the means are equal, there should be only a small probability that when applying the test the statistical inference leads to the assertion that the means are not equal. The second requirement states, that if a difference does exist between the two means, there should be a very high probability that this fact is detected by the test. Sverdrup (26) points out that in effect these two requirements are competing with one another, and in choosing any test of hypothesis both considerations must be balanced against one another. On one hand there is a strong desire to claim that the two means are equal when in fact they are equal. However, at the same time an equally strong desire exists which concentrates on detecting the smallest possible difference between the two means in an attempt to assert that the two means are not equal when they are not equal. If the first requirement is too strongly adhered to then the probability of detecting a difference between the means when it exists is decreased, thereby weakening the second requirement. Conversely, when the test attempts to detect

extremely small differences between the two populations means, the probability of asserting that the means are not equal, when in fact they are equal, will increase.

In hypothesis testing a statement whose erroneous rejection it is particularly desirable to avoid, is called the null hypothesis, and is generally denoted by H_o. In the case of the t-test the null hypothesis is therefore the statement that the means of the two populations are equal. If the means are equal it is not desirable to conclude from statistical inference that they are not equal. If the means are truly not equal it is not desirable to conclude that they are after using the test. This situation is schematically shown in Table 1.

Table 1
ERRORS IN HYPOTHESIS TESTING

		TRUE SITUATION	
		NULL HYPOTHESIS TRUE	NULL HYPOTHESIS FALSE
TEST INDICATES	ACCEPT NULL HYPOTHESIS	NO ERROR	TYPE II ERROR
	REJECT NULL HYPOTHESIS	TYPE I ERROR	NO ERROR

A Type I error results when the null hypothesis is rejected when in fact it is true and a Type II error results when the null hypothesis is accepted when in fact it is false. Symbolically the probability of making a Type I error is denoted by α and the probability of committing a Type II error is denoted by β. The probabilities associated with making a Type I or Type II error should be as small as possible.

The critical importance in understanding these two criteria is the fact that they will be the basis of the evaluation for the t-test during this study. When two populations meet all three of the assumptions necessary for use of the t-test, the test results in a certain fraction of Type I and Type II errors which are unavoidable. This investigation examines in detail how these fractions change when the assumption of equal variances is violated.

B. SIGNIFICANCE LEVEL

The tabulated t value mentioned earlier will now be referred to as the critical t value or t_{crit}. The particular value of t_{crit} is choosen such that a fraction α of the distributional values of the t distribution lie beyond $|t_{crit}|$. This is the result of having the null hypothesis $H_o: \mu_x = \mu_y$ and choosing the alternative hypothesis $H_1: \mu_x \neq \mu_y$. That fraction of the distributional values lying outside of $|t_{crit}|$ is equal to α , the probability associated with a Type I error.

If the two population means are equal and the t value resulting from the t-test lies outside of the interval $(-t_{crit}, t_{crit})$, the test produces a Type I error. This is due entirely to chance with a probability equal to α and this type I error is unavoidable in an α fraction of the cases run.

The signficance level of the test is equal to one minus the probability of making a Type I error and is written symbolically as $1-\alpha$.

C. POWER OF A TEST

The probability of committing a Type II error is denoted by β. This is the proportion of acceptances of the null hypothesis when in fact the hypothesis should be rejected. The power of any test is defined as $1-\beta$. As β increases the power decreases and conversely as β decreases the power of the test increases. It results that when two normal population means are almost equal the power of the test is small and the power increases as the difference in the means increases. As the difference between the means does increase the power of the test asympototically approaches 1.0. When no difference exists between the population means then β equals $1-\alpha$.

The power of any statistical test is a function of certain factors. The principle factor influencing the power is the variance of the respective populations being tested.

The test being evaluated could be influenced by the largest variance of the two populations, the magnitude of the difference between the two population variances or the size of the pooled variance for both populations. A second factor influencing the power of a test is the size of the samples taken from both populations and whether or not these sample sizes are equal. The sample sizes have a strong influence on the size of the pooled variance. The pooled variance (pv) is defined as

$$pv = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2} \qquad 2\text{-}1$$

When the sample sizes are equal the pooled variance is simply one-half of the sum of the variances from both populations. When the sample sizes are not equal then the size of the pooled variance is most effected by the sample having the larger number of observations.

III. VIOLATION OF ASSUMPTIONS

A. PREVIOUS INVESTIGATIONS

Very few investigations have been carried out to study the effect of dependent random samples on the Student's t-test. Scheffe´ (25) discusses a violation of this nature and proves that the effect of a serial correlation on inference about means can be serious and, therefore, should be considered when using the test. With respect to the normality assumption it is usually reasonable to assume normally distributed populations because even when populations are not normal Scheffe´ (25) has demonstrated that the effect of a violation of this nature is very slight when making inferences about means.

The most interesting and most complex results arise when the assumption of equal variances is violated. Circumstances often exist where group to group homogeneity of variances is not to be expected and is the exception rather than the rule.

For the particular case where non-homogeneity of variances is known to exist, different methods have been proposed as alternatives to the t-test. When the relative scale factor of the two populations is known appropriate weighting of the sums of squares gives an exact solution. In the case where the relative scale factor is unknown different criteria have been advocated.

Welch (30) has discussed in detail the often employed alternative statistic

$$w = \frac{\bar{x} - \bar{y}}{\left[\frac{\sum_{i=1}^{n_x} (X_i - \bar{X})}{n_x(n_x - 1)} + \frac{\sum_{i=1}^{n_y} (Y_i - \bar{Y})}{n_y(n_y - 1)} \right]^{\frac{1}{2}}} \qquad 3\text{-}1$$

He demonstrates that when $\alpha_x^2 \neq \alpha_y^2$ the t statistic developed in 1-6 does not have an underlying t distribution and that 3-1 results in less bias than the general t statistic when the variances are not equal.

Fisher (5) has proposed another solution to the problem of testing the hypothesis $\mu_x = \mu_y$ using the concept of fiducial distributions but the validity of this approach has been questioned by Bartlett (1).

Each of these alternatives was developed because the contention exists that the t-test is not generally applicable to testing the equivalence of means when the variances of the two populations may not be equal. This study is not concerned with comparing these alternatives with the t-test, it will attempt to determine the necessity of using these alternatives. The t-test may prove to be robust enough to withstand such a high degree of violation of the assumption of equal variances that these alternatives are not necessary.

Welch (29) made the first detailed study of the t-test and its robustness when faced with a violation of the assumption of equal variances. He concentrated on only the resulting α level and used an approximation method to arrive at his results. When the sample sizes were equal, Welch's conclusion was that the rejection rate arrived at when the variances are different does not differ significantly from the specified rate. The approximation used, set the variation of one population to zero and even under this extreme condition the test never became seriously biased. In terms of frequencies, Welch has stated that for equal sample sizes and a difference in population variances, if the test were performed numerous times the number of rejections of a true hypothesis would not be significantly different than the actual number of expected rejections for a prescribed α level. Using the t-test as an example, if the test were applied many times to two normal populations with equal means, the number of Type I errors expected would be equal to the fraction α of the total number of iterations of the test. If the two population variances were in fact different, approximately this same number of expected Type I errors would result. Therefore, the violation of the assumption of equal variances does not bias the test seriously when the sample sizes are equal. This investigation attempts to verify empirically the truth of Welch's statements.

Welch also examined the case where the sample sizes were not equal. Using the same approximation method he made the following observations. When the larger sample has the larger variance the difference between the two means tends to be underestimated. This implies that the probability of making a Type II error increases, and consequently the power of the test will decrease. When the larger sample has a smaller variance the difference between the two means tends to be overestimated and a greater percentage of Type I errors result. The foregoing result could be summarized to state that the true rejection rates becomes significantly different than the specified rates for unequal sample sizes and unequal population variances.

Gronow (9) likewise made an exhaustive study of the rejection rate of the t-test when the assumption of equal variances is violated. He used a different method of approximation then Welch, but his study resulted in confirming what Welch had previously stated. A bias will result in the rejection rate for populations with unequal variances and different sample sizes.

In both of these previous investigations, Welch and Gronow were hampered by the fact that they had to use an approximation method to arrive at their conclusions. Consequently, they were forced to look at extreme cases and

draw conclusions. The ratio of variances was set either at 0, 1 or ∞, and then through a mathematical argument they arrived at a result. This approach leaves many fine points unanswered. For instance, Welch used equal sample sizes of ten observations each and made his conclusions concerning the lack of bias with respect to rejection rates. The question of what happens with rejection rates for equal but smaller sizes remains unanswered. Is there a variance ratio large enough to cause the "true" rejection rate to differ significantly from the specified rate? For the same reason the use of extreme cases did not yield enough information to draw definitive conclusions concerning the power of the t-test under varying variance ratios.

The rapid development of high speed computers within the last ten years has been largely responsible for making detailed studies in this area more feasible. Murphy (19) used computer simulation to test the actual rejection rates while comparing the t-test to two alternatives, the Permutation Test and the Aspin-Welch Test. At a specified α level of 0.05 he substantiated Welch's and Gronow's work concerning the bias inherent in the test when the sample sizes differ and population variances are not equal. During his investigation, Murphy used 500 iterations for each case studied.

B. AREAS OF INVESTIGATION

These previous investigations into the characteristics of the t-test aid and encourage further study. The mathematical results furnished by Welch and Gronow beg for substantiating data in the form of numerous applications of the t-test under various degrees of violation of the assumption of equal variances. This investigation attempts to provide this needed data while it studies the effect of unequal variances on the robustness of the test. It should be restated that robustness of a test is concerned with the fraction of Type I and Type II errors exhibited by the test. A study of Type I and Type II error rates and the power of the test determines the effect of this violation on robustness.

The rejection rates of the test are studied for varying degrees of unequal variances. The ratio of the two population variances is termed the scale factor k, and this scale factor is allowed to range over intervals determined from the investigation. With equal and unequal sample sizes an attempt is made to find the particular value of k, if one exists, where the actual or estimated "true" rejection rate differs significantly from the specified α level of 0.05. A second method for finding a particular k value is used. An accumulation of observations are made for certain other α levels and combining these

figures results in the formation of the tail of an empirical frequency distribution which is compared to the tail of the theoretical t distribution to determine if the violation of the assumption of homogeneity of variance causes the t-test to produce an empirical distribution which differs significantly from the t distribution. Once again the attempt is made to find a particular value of k which marks a point where the empirical frequency distribution no longer parallels the t distribution.

The investigation attempts to substantiate Welch's conclusion that for unequal samples the t-test quickly becomes invalid under the violation of the assumption of equal variances, or to show that the validity of the test is only violated at such an extreme scale factor that in effect the test is valid in most circumstances. A test is valid if it functions as intended with respect to the two criteria in hypothesis testing. This means that the values of α and β are the primary measures of effectiveness for this investigation.

The power of the test is also investigated in the cases of equal and unequal sample sizes. It is desirable to determine if the power of the test decreases as the scale factor varies from k=1, and further, if the power does decrease, is the change due to the violation of the assumption of equal variances or is the decrease in some way related to the actual variance present in both samples?

IV. METHODS AND PROCEDURES

A. METHODS

Computer simulation was used to carry out the investigation. The investigation took the form of programming numerous "cases" through the computer. Each case, which was iterated 50,000 times, consisted of the following elements:

1. Two samples drawn from each of two standard normal populations, X and Y. The sample sizes were n_X and n_Y, and ranged in size from five to fifteen observations each and were not always equal.

2. A scale factor k equal to the ratio of variances, σ_X^2 / σ_Y^2 where k was allowed to vary discretely over a determined range. The values of the variances from the two normal standard populations, N(0,1) were adjusted to achieve the desired scale factor.

3. A difference in means of the two populations which was allowed to range from zero to five, in 0.5 increments, which resulted in 11 different values.

As an example, a single case would consist of $n_X = 10$, $n_Y = 8$, $k = 5$ and $\mu_X - \mu_Y = 3.5$. For this case, 50,000 iterations were performed and the following data were gathered: the rejection rates for the critical values of the t distribution associated with α levels of 0.1,

0.05, 0.02, 0.01, and 0.001 were compiled. At the level of 0.05, the estimate of the "true" rejection rate α_t and the estimate of the "true" power of the test $1-\beta_t$ were calculated .

Initially, 5,000 iterations were performed for each case. This was done to arrive at some indication of what value the scale factor had to obtain to force the test to produce invalid inferences. When this tentative scale factor was determined for each pair of sample sizes the number of iterations was increased to 50,000 and the scale factor was allowed to vary from one to this tentative value in increments of 0.25.

Two different criteria were used to determine the "validity" of the t-test at various scale factor or k values. First a study was made of the differences between the estimated "true" rejection rate resulting from the 50,000 iterations and the expected rejection rate at a single α level of 0.05. These two rejection rates were compared to determine at what k value they became significantly different. The test used to conduct this comparison had a significance level of 0.975.

The second method used to determine the "validity" of the t-test was more stringent then the comparison of rejection rates at a single α level. The second method took the rejection rates compiled at the five α levels,

0.10, 0.05, 0.02, 0.01, and 0.001 and from these figures constructed the tail of an empirical frequency distribution. This developed empirical distribution was then compared to the tail of the t distribution to determine at what k value the two distributions became significantly different. A Chi Square Goodness-of-Fit Test with four degrees of freedom and a significance level of 0.975 was used to conduct the comparison.

Also during the 50,000 iterations for each case the estimated "true" rejection rate for Type II errors was being compiled and converted into a value for the power of the test. Appropriate cases were combined to develop power curves for graphic comparisons.

B. PROCEDURES USED

Sample generation was accomplished with a Gaussian Normal Generation Program on file with the computer center at the Naval Postgraduate School. The program was developed by Marsaglia, MacLaren, and Bray (15). The authors stated that in theory the Gaussian method they developed is completely accurate in that the procedure employed returned a random variable with exactly the required distribution, and in practice the result is an approximation influenced only by the capacity (word length) of the computer used.

The accuracy of the random variables generated was tested by studying the first four moments, mean, standard deviation, skewness, and kurtosis on 35 samples of 10,000 numbers each. Each sample generated a distribution with normal characteristics. A χ^2 Goodness-of-Fit Test with nine degrees of freedom and a 0.99 significance level was also used to test the 35 samples. Using this test the samples were tested against a N(0,1) population and no significant differences resulted between any of the samples and this N(0,1) population. These investigations seemed to give adequate indication that the numbers being generated were from N(0,1) population.

The actual method of obtaining the information called for in the study consisted of using the FORTRAN Program included in Appendix A. In the program the sizes of the two samples were initially established. Sample sizes ranged from five to fifteen observations and n_X and n_Y could be set to any value within the range. Initially both samples were drawn from a N(0,1) population using the Gaussian Normal Generation Program. By multiplying each observation of one sample by a standard deviation value σ, and adding a constant, c, to the result, the underlying population of the sample was transformed into a desired normal population, $N(c,\sigma^2)$. The two normals, N(0,1) and $N(c,\sigma^2)$ now had a variance ratio of $1/\sigma^2$ and a difference

in means equal to c. The two samples were then subjected to the t-test and the resulting t statistic was tabulated for the appropriate rejection rates. This iteration was cycled 50,000 times. At the conclusion of the iterations the value for the difference in means was incremented, the standard deviation value remained the same, and another case with 50,000 iterations was performed. When all values for the differences in means had been exhausted, a new value for the standard deviation was read into the program and the entire process repeated. This procedure was continued until all desired variance ratios were generated.

Tabulation of the rejection rates consisted of testing the resulting t statistic against appropriate critical values. The particular critical values chosen were not only a function of the desired α level but also the number of degrees of freedom for the particular case. The degrees of freedom for any case were equal to the total number of observations from both samples minus two, (i.e., $n_x + n_y - 2$). This number of degrees of freedom results from the fact that there are $n_x - 1$ independent deviations from the mean in the first sample and $n_y - 1$ in the second and a total of $n_x + n_y - 2$ independent deviations from the mean to estimate the populations' variances.

V. RESULTS

A. ESTIMATED "TRUE" REJECTION RATES

1. Equal Sample Sizes

The initial objective in this study was to investigate what effect a violation of the assumption of homogeneity of variances would have on the rejection rate of the t-test, at $\alpha = 0.05$. At what k value would the estimated "true" rejection rate differ significantly from the expected rejection rate?

Initially the cases for equal sample sizes were studied. Samples of size five, ten, and fifteen were chosen. It was assumed that information gathered at these levels would cover the complete spectrum of possible results encountered in the use of the t-test. Table 2 below gives the results of the estimated "true" rejection rates of the t-test over the range of scale factors, when samples of equal sizes were used.

Table 2
ESTIMATED "TRUE" REJECTION RATES FOR $\alpha = 0.05$,
EQUAL SAMPLE SIZES

k		1/9	1/7	1/5	1/3	1	3	5	7	9
n_x	n_y									
5	5	.0686	.0656	.0564	.0556	.0494	.0542	.0600	.0662	.06
10	10	.0578	.0536	.0512	.0540	.0440	.0474	.0530	.0616	.06
15	15	.0558	.0554	.0554	.0532	.0486	.0514	.0536	.0486	.05

The values given in Table 2 are the fraction of rejections of 5,000 iterations in each case. With an α level of 0.05, the expected rejection rate is exactly 0.05. Even in the cases where all the assumptions are completely satisfied the expected rejection rate can only closely approximate 0.05 because the number of rejections is a random variable from a binomial distribution with parameter p = .05. The occurence of a rare event has positive probability and therefore small deviations from 0.05 can occur for the expected rejection rate. It can be seen that as k deviated from one in both directions, the estimated "true" rejection rate also increased with respect to the α level of 0.05. This occurence was true for each of the equal sample sizes. As the sample sizes themselves increased and more information was available to the t-test, there seemed to be a less rapid growth in the difference between the "true" and specified rejection rates.

The k values in Table 2 were developed by setting the variance of the Y population equal to one and then allowing the variance of the X population to change in order to effect the desired variance ratio. This meant that even for equal sample sizes k values of k = 1/9 and k = 9 were not exactly the same. For both scale factors the magnitude of the ratios of the two population variances is the same but the pooled variance present in case k = 1/9

is 5/9 and in the case k = 9 the pooled variance is 5. This same type of difference is present in other complimentary pairs of k values, 1/3 - 3, 1/5 - 5, and 1/7 - 7. In observing the data though there appears to be no correlation between the size of the pooled variance and a change in the estimated "true" rejection rate. It was concluded that the primary cause for a change in the estimated "true" rejection rate was a change in the scale factor value.

The primary objective of the investigation was to determine those values of k at which the estimated "true" rejection rate begins to differ significantly from the specified α level. A Chi Square test with one degree of freedom and a significance level of 0.975 was used to determine the fraction, and number of "true" rejections that if achieved by the test, would imply that the two rates could be considered significantly different. The χ^2 statistic was developed from the case shown below.

	NUMBER OF CASES REJECTED	NUMBER ACCEPTED
OBSERVED	A	B
EXPECTED	250	4750

The expected number of rejections, 250 comes from the fact that 5,000 iterations were performed for each case and the critical t value used produced a specified α level of 0.05. Five percent of 5,000 is 250, the expected number of rejections.

Using a 0.975 significance level for the χ^2 test meant that if the number of observed rejections, A, became greater than 319 or less than 181, a significance difference between the estimated and specified rejection rates would be implied. Three hundred and nineteen is exactly 6.38 percent of 5,000, and 215 is exactly 3.62 percent of 5,000.

With these critical percentages of .0638 and .0362 and the data from Table 2, the following observations can be made. For the sample sizes of five observations the critical value of k, where the estimated "true" rejection rate becomes significantly different from the specified rate appears to occur for a k value between five and seven. For equal sample sizes of either 10 or 15 observations each the sought after critical k value appeared to lie beyond k = 9. It was decided to conduct the investigation for these two equal sample sizes for k values between one and nine.

The more detailed study was now conducted. For equal sample sizes of 5, 10, and 15 observations the k intervals (1,5), (1,9), and (1,9) respectively were

investigated. In each case the variance ratio was incremented from one to the upper limit of the interval in 0.25 steps. At each scale factor value 50,000 iterations were performed. For 50,000 iterations and an α equal to 0.05, the critical number of rejections became either 2718 or 2282. For any k value producing a number of rejections greater or less than these two figures respectively, the implication would result that the estimated "true" rejection rate was significantly different from the expected rejection rate.

At the same time the 50,000 iterations produced rejection rates for the other specified α levels, 0.10, 0.02, 0.01, 0.001. With these rates it was possible to develop an empirical frequency distribution. By comparing this empirical distribution with the t distribution it was possible to determine, in a second manner, a critical k value where the two distributions became significantly different.

The results of using these two criteria for testing the validity of the t-test for the various equal sample sizes under varying k values is contained in Table 3. The k values listed include all the pertainent information needed in the investigation.

Table 3
VALIDITY RESULTS FOR THE t-TEST WITH EQUAL SAMPLE SIZES.
50,000 ITERATIONS AT EACH k VALUE

$n_x = n_y =$	5 Criteria		10 Criteria		15 Criteria	
k	A	B	A	B	A	B
1.00	2502	A	2455	A	2420	A
1.25	2434	A	2503	A	2490	A
1.50	2589	A	2545	A	2484	A
1.75	2526	A	2514	A	2425	A
2.00	2588	A	2562	A	2656	A
2.25	2614	A	2571	A	2490	A
2.50	2679	R	2564	A	2569	R
2.75	2737	R	2576	A	2537	A
3.00	2819	R	2597	A	2545	A
3.25	2758	R	2650	R	2572	A
3.50	2917	R	2730	R	2627	R
3.75	2904	R	2716	R	2575	A
4.00	2887	R	2706	R	2774	R
4.25	2946	R	2686	R	2580	R
4.50	2954	R	2726	R	2693	R
4.75	3030	R	2722	R	2640	R
5.00	3179	R	2745	R	2671	R
5.25			2779	R	2671	R
5.50			2845	R	2693	R
5.75					2651	R
6.00					2860	R
6.25					2688	R
6.50					2773	R
6.75					2683	R
7.00					2734	R
7.25					2708	R
7.50					2752	R
7.75					2726	R
8.00					2881	R
8.25					2730	R

A - Estimated "true" number of rejections at single α level of 0.05, critical number 2718 or 2282 (α= 0.025)

B - Outcome of testing H_o that the empirical distribution equals the t distribution (α = 0.025)

In each of the cases of equal sample sizes, as the scale factor k, increased the estimated number of "true" rejections for an α level of 0.05 also increased. For equal sample sizes, five observations each a definite k critical value between 2.50 and 2.75 was determined where the estimated "true" rejection rate differed significantly from the expected rejection rate of 2,500 rejections in 50,000 iterations. For samples of ten observations each such a definitive break is not so evident. At k = 3.50 the two rejection rates are significantly different while for k = 3.75, 4.00, and 4.25 the rates are not significantly different. For k values greater than 4.25 the two rates are consistently significantly different. The assumption of the result at k = 3.50 is an extreme random occurrence, results in concluding that the estimated "true" rejection rate begins to differ significantly from the expected rejection rate at a scale factor of k between 4.25 and 4.50. Such a random occurrence is also assumed to have occurred in the case of 15 observations each and k = 4.00. This particular case yielded rather inconclusive results and it can only be determined that the critical k value sought for lies in the k range from 5.75 to 6.75.

The results of using this less stringent requirement can be summarized in Table 4.

Table 4
CRITICAL k INTERVALS DETERMINED UNDER THE CRITERIA OF EQUAL REJECTION RATES

EQUAL SAMPLE SIZES n=	CRITICAL INTERVAL k*
5	2.50-2.75
10	4.25-4.50
15	5.75-6.75

In evaluating the robustness of the t-test with respect to a significant difference between the developed empirical distribution and the t distribution the resulting critical k intervals determined were less in all cases than the k intervals discussed in the previous paragraph. For the case $n_x = n_y = 5$, the k value where the two distributions became significantly different occurred in the interval 2.25 to 2.50. In the case $n_x = n_y = 10$, the hypothesis that the two distributions were equal was accepted up to a k value between 3.00 and 3.25. A variance ratio greater than 3.25 produced a rejection of the hypothesis without exception. In the case $n_x = n_y = 15$ such an exact k interval could not be determined. Rejections of the hypothesis occurred at k equal to 2.50, 3,50, and values greater than or equal to 4.00. Assuming that this case is as robust as the case for ten observations in each sample, the rejection at k = 2.50 could be considered an extreme random occurrence. Because of the rejection of the hypothesis at k = 3.50 no concise 0.25 k interval

appears to exist. Therefore it was only concluded that the critical k value sought after must lie in the interval between k = 3.25 and k = 4.00.

The results of using this more stringent requirement are summarized in Table 5 below.

Table 5
CRITICAL k INTERVALS DETERMINED UNDER THE CRITERIA OF EQUAL DISTRIBUTIONS

EQUAL SAMPLE SIZES n=	CRITICAL INTERVAL k*
5	2.25-2.50
10	3.00-3.25
15	3.25-4.00

Even for the most stringent criteria and the smallest equal sample sizes, five observations, the k* found was between 2.25 and 2.50. This means that the variances of the two normal distributions under study can differ in magnitude by a factor greater than two and the t-test can still give valid answers. Increasing the observations to 15 in each sample allows the variances to differ in magnitude by a factor of approximately four, and the t-test still continues to produce valid inferences. Reducing the stringency of the criteria for validity increases the degree of violation of the assumption that the t-test can withstand. With respect to estimated "true" rejection rate, and equal sample sizes this segment of the investigation indicates that the t-test is extremely robust.

2. Unequal Sample Sizes

Welch (29) had predicted that for unequal sample sizes a violation of the homogeneity of variance assumption would result in a strong bias and invalidate the t-test rapidly. Unequal sample sizes were studied in the same manner as the equal sample size cases. Initially 5,000 iterations were performed to obtain an indication of what range of k values were needed to be included in a more detailed study. These initial results are contained in Table 6.

Table 6
ESTIMATED "TRUE" REJECTION RATES FOR $\alpha = 0.05$, UNEQUAL SAMPLE SIZES, $\alpha_y^2=1$

n_x	n_y	k = 1/9	1/7	1/5	1/3	1	3	5	7	9
8	6	.0944	.0924	.0870	.0748	.0498	.0432	.0378	.0398	.0432
10	6	.1226	.1116	.1056	.0844	.0504	.0290	.0240	.0242	.0218
13	6	.1652	.1586	.1270	.1074	.0462	.0180	.0154	.0140	.0122
15	6	.1898	.1764	.1634	.1126	.0526	.0162	.0114	.0082	.0074
15	10	.0968	.1020	.0930	.0782	.0512	.0330	.0294	.0274	.0280
15	12	.0840	.0736	.0708	.0646	.0482	.0356	.0364	.0392	.0366
11	8	.0944	.0936	.0826	.0710	.0490	.0324	.0360	.0350	.0324

The bias characteristic of the test is evident from the data of Table 6. Remembering that k, the scale factor, is defined as α_x^2/α_y^2, the table shows that whenever the larger sample n_x has the larger variance, k = 3, 5, 7, or 9, the estimated "true" rejection rate is less than the specified rate. When the sample n_x has the smaller variance, k = 1/3,

1/5, 1/7, or 1/9, the estimated "true" rejection rate is greater than the specified rate. This observation is true in all cases and is an actual data confirmation of Welch's mathematical conclusions.

To explain this result, the formula for the t statistic must be further examined where

$$t = \frac{\bar{x} - \bar{y}}{\left[\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}\right]^{\frac{1}{2}} \left[\frac{1}{n_x} + \frac{1}{n_y}\right]^{\frac{1}{2}}}$$

Of importance is the first term of the denominator. This quantity is called the pooled variance and is the critical term in explaining the results in Table 6. To obtain the desired scale factor k the variance for the Y population was maintained at one and the variance for the X population was allowed to vary to achieve the particular scale factor. For any of the unequal sample cases in Table 6 with k = 1, the pooled variance term of the t statistic came out to a certain average result. Now as k increased from one through nine the sample variance of the X population, s_x^2, also increased. This caused the pooled variance term to also increase and with the remaining term of the denominator and the numerator remaining relatively constant the average t statistic decreased. As the t statistic decreased a greater proportion of the results fell within the critical

interval $(-t_{crit}, t_{crit})$, and the probability of a t statistic greater than t critical decreased. The estimated "true" rejection rate therefore decreased. In an opposite manner, as s_x^2 decreased, k = 1 to 1/9, the average t statistic increased and a greater proportion of the results fell outside of the critical interval causing the estimated "true" rejection rate to increase.

In the pooled term the sample variances is weighted by $(n_x - 1)$. Now for any particular k value, as n_x increases the change in the estimated "true" rejection rate is accelerated. As an example, for k = 3, in all the cases where $n_y = 6$ the estimated "true" rejection rate is less than the specified rate. Proceeding down the column, as n_x increases the difference between the two rates is increasingly more pronounced. This is due to the increased weight applied to s_x^2 as n_x increases.

This same bias was investigated by developing the scale factor k by a different method. In this instance the variance for the X population was set equal to one and the variance of the Y population was allowed to vary in order to develop the desired scale factor values. The same type of bias characteristics were obtained and are shown in Table 7. In a majority of the data points the bias was slightly more pronounced in each direction when compared to similar points in Table 6 but they do not

appear to be significantly different. When the bias caused the estimated "true" rejection rate to be greater than the specified level the bias was even greater in the cases where $\alpha_x^2 = 1$. This difference, though slight, between the two approaches can be explained. In Table 7 the smaller sample size n_y is drawn from the population with the changing variance. Statistically, this smaller sample provides less information about the underlying population, with the resulting mean standard deviation being greater than the case where the sample variance of the larger sample is varied, thus the bias is more pronounced.

Table 7
ESTIMATED "TRUE" REJECTION RATES FOR $\alpha = 0.05$,
UNEQUAL SAMPLE SIZES, $\alpha_x^2 = 1$

		k								
n_x	n_y	1/9	1/7	1/5	1/3	1	3	5	7	9
8	6	.1038	.0886	.0864	.0788	.0498	.0446	.0380	.0388	.0420
10	6	.1312	.1172	.1142	.0854	.0504	.0310	.0258	.0236	.0252
13	6	.1608	.1556	.1340	.1070	.0462	.0212	.0130	.0146	.0114
15	6	.1818	.1866	.1542	.1198	.0526	.0172	.0112	.0078	.0086
15	10	.1028	.1042	.0976	.0786	.0512	.0312	.0284	.0324	.0232

The explaination of the bias characteristics discussed for the case resulting in Table 6 also applies for the method of generating the scale factor in this case. The same results hold in that the greater the difference between sample sizes the more pronounced the bias.

In searching for a critical k value in each of the unequal sample size cases, the initial 5,000 iteration test revealed that in every case except for $n_x = 8$, $n_y = 6$, the estimated "true" rejection rate became significantly different from the expected rejection rate at k values less than 3.00. Therefore the initial k values tested for 50,000 iterations ranged over the interval from 1/3 to 3. If any case indicated a critical k value existed outside of this interval then the range could be increased. From the results contained in Table 8 it is evident that no increase in the k range was necessary for any of the cases studied.

Table 8
VALIDITY RESULTS FOR THE t-TEST WITH UNEQUAL SAMPLE SIZES, 50,000 ITERATIONS AT EACH k VALUE

n_x-n_y	8-6 Criteria		10-6 Criteria		13-6 Criteria		15-6 Criteria		15-10 Criteria		15-12 Criteria	
k	A	B	A	B	A	B	A	B	A	B	A	B
0.333	3661	R	4428	R	5401	R	5926	R	3963	R	3312	R
0.364	3516	R	4261	R	5113	R	5547	R	3851	R	3321	R
0.400	3402	R	3949	R	4870	R	5278	R	3660	R	3214	R
0.444	3250	R	3872	R	4434	R	4946	R	3455	R	3063	R
0.500	3110	R	3672	R	4076	R	4460	R	3350	R	3004	R
0.571	2997	R	3392	R	3874	R	4110	R	3161	R	2789	R
0.666	2811	R	3060	R	3477	R	3591	R	2992	R	2825	R
0.800	2726	R	2766	R	2958	R	3034	R	2785	R	2633	R
1.000	2501	A	2524	A	2418	A	2481	A	2392	A	2440	A
1.250	2411	R	2196	R	2052	R	2016	R	2302	R	2308	R
1.500	2168	R	2020	R	1852	R	1722	R	2108	R	2321	R
1.750	2173	R	1872	R	1504	R	1448	R	1958	R	2160	R
2.000	2105	R	1792	R	1436	R	1214	R	1883	R	2088	R

n_x-n_y	8-6 Criteria		10-6 Criteria		13-6 Criteria		15-6 Criteria		15-10 Criteria		15-12 Criteria	
k	A	B	A	B	A	B	A	B	A	B	A	B
2.25	2037	R	1644	R	1289	R	1163	R	1769	R	2094	R
2.50	2016	R	1563	R	1145	R	1014	R	1681	R	2146	R
2.75	1995	R	1575	R	1083	R	922	R	1731	R	2115	R
3.00	2008	R	1490	R	1032	R	854		1622	R	2069	R

A - Estimated "true" number of rejections at single α level of 0.05, critical number 2718 or 2282 (α= 0.025)

B - Outcome of testing H_o that the empirical distribution equals the t distribution (α= 0.025)

Table 8 continued

Using either criteria for testing the validity of the t-test for different k values the results indicated that for unequal sample sizes the robustness of the t-test is poor. For every case the slight increase in k to a value of 1.25 caused a violation of the criteria that the developed empirical distribution and the t distribution must not be significantly different. The less restrictive criteria that the estimated and expected rejection rates be equal was violated at k value very close to one. Only in the case n_x = 15, n_y = 12 could a k value in the range 1.25 to 1.50 be tolerated by the test.

These results demonstrate rather emphatically Welch's predictions that for unequal sample sizes a violation of the homogeneity of variance assumption would result in a strong bias and invalidate the t-test rapidly. The t-test was not able to withstand a violation of the assumption to any degree and the robustness of the test in this instance must be considered extremely poor.

B. POWER

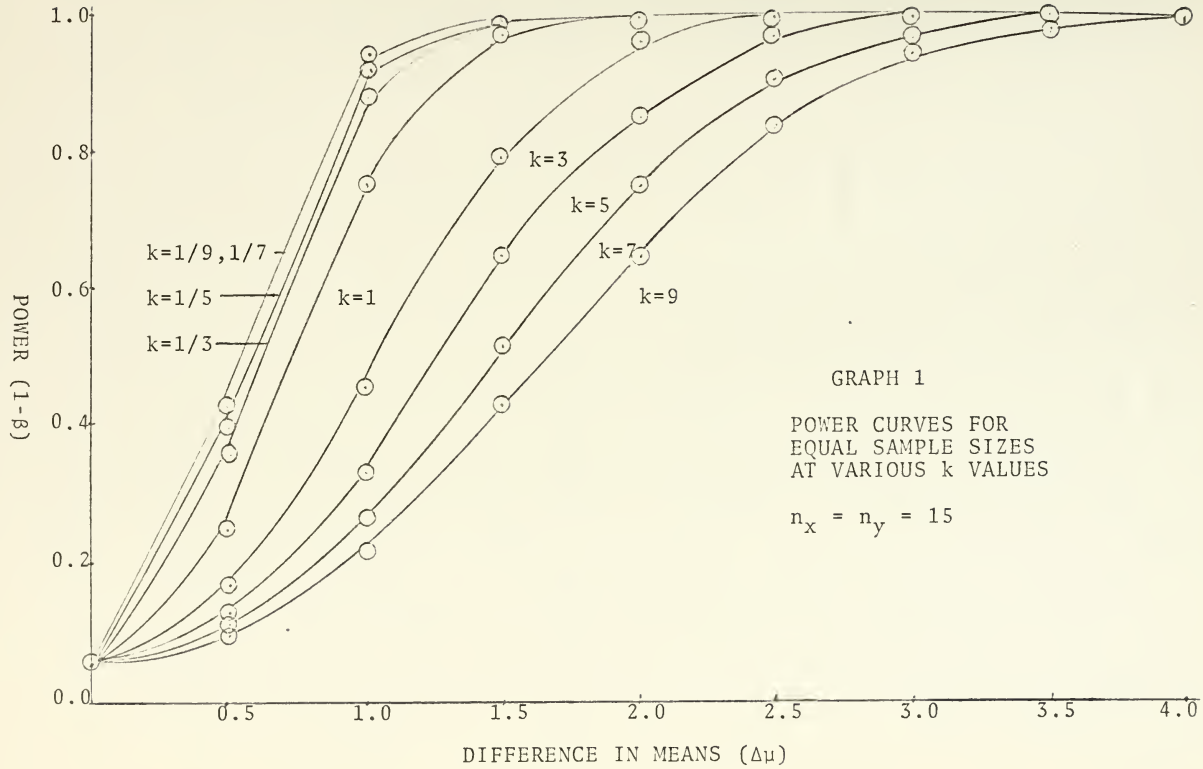
The power of the t-test was investigated in a similar method as the Type I error rate. Cases were studied for both equal and unequal sample sizes and various degrees of violation of the assumption of equal variances. The Type II error (β) of accepting the null hypothesis when in fact it should be rejected because the populations means are not equal was used to develop the power of the t-test, $1-\beta$ and conclusions were made through comparisons of graphic results. In all cases an α level of 0.05 was used.

The primary question asked in the investigation was what effect did a violation of the equal variance assumption have on the power of the test? Was a change in the power directly related to the degree of the violation or did there exist a more important factor in determining the power of the test? As discussed in Chapter 2 the power of any test is influenced by a combination of factors, variances, and sample sizes.

1. Equal Sample Sizes

The results illustrated in Graph 1 are for equal sample sizes, 15 observations each and are typical of each of the other equal sample size cases of five and ten observations. Data gathered for each of these cases are contained in Table 9. Graph 1 indicates that as k increased in value from one to nine the power of the test decreased. This is a predictable result because of the increased variance present in the X population. Also shown though in Graph 1 is the result that as k decreased from 1 to 1/9 the power of the test increased. To explain this result it should be remembered that the desired k values were achieved by maintaining α_y^2 constant and equal to 1 and programming α_x^2 equal to specific values. This means that as k increased from 1/9 to 9 the pooled variance (2-1) also increased, and as can be seen the power of the test decreased. In the range from k = 1/9 to k = 1 there was a relatively small decrease in the power but this is explained by the fact that the variance of the X population had to increase in relatively small increments to achieve the desired k values. Therefore in this range the size of the pooled variance increased only slightly.

Power decreased appreciably in the range k = 1 to k = 9 because of the relatively large increases in the variance of the X population. The pooled variance also exhibits this relatively large increase over this same range of k values.



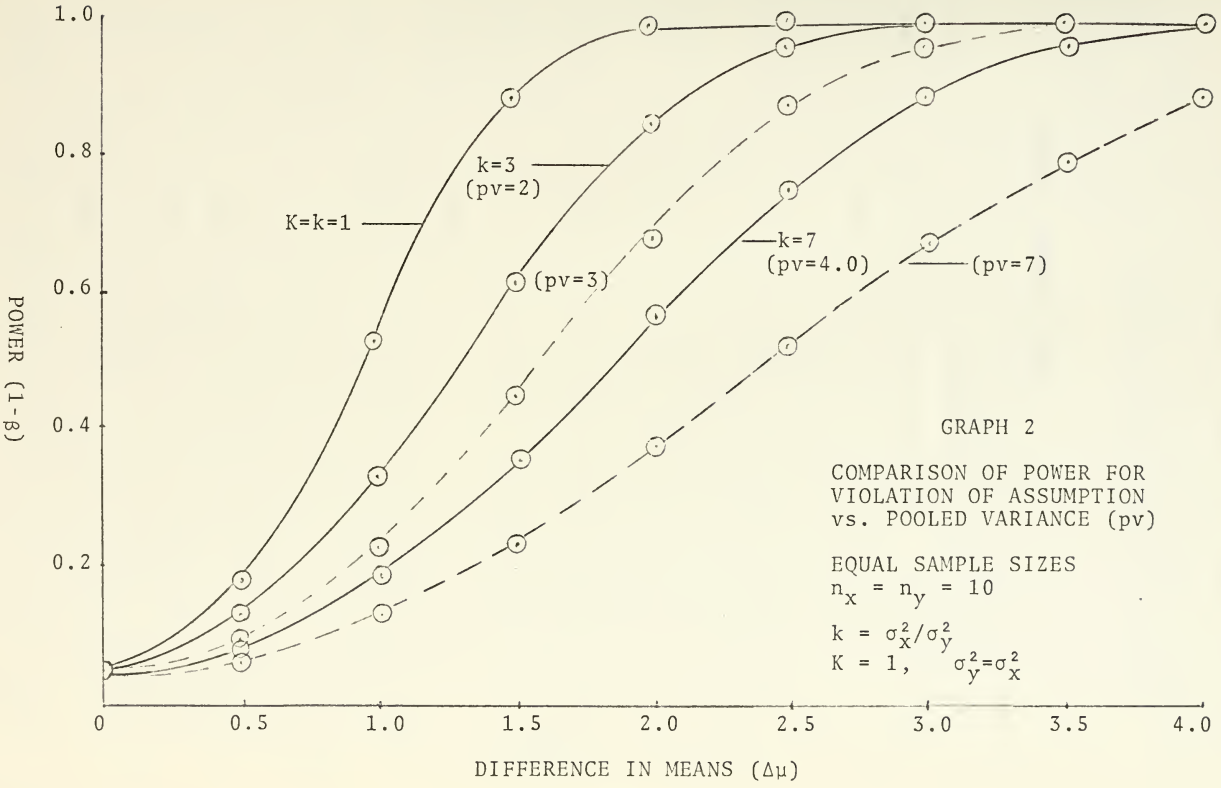
GRAPH 1

POWER CURVES FOR
EQUAL SAMPLE SIZES
AT VARIOUS k VALUES

$n_x = n_y = 15$

The conclusion made from these observations is that a violation of the assumption of equal variances does not directly influence the power of the t-test. There is a significant difference in the power for k=1/9 and k=9 even though the degree of violation is the same in both cases. The power of the test is directly a function of the size of the pooled variance and the less the amount of pooled variance the greater the power of the test.

To emphasize the contention that the size of the pooled variance is the primary factor influencing the power of the t-test, Graph 2 is provided. Two sets of curves are plotted. There are two curves with scale factors equal to 3 and 7 and they are compared to two curves (K) where the scale factor is equal to one and therefore no violation of the assumption exists but the size of both population variances are equal to 3 or 7. For k=3 the pooled variance is equal to 2. For K=1, $\alpha_y^2=\alpha_x^2=3$ the pooled variance is equal to 3. The power of the k=3 curve is greater than for K=1 and the variances equal to 3, but this same curve (K=1) exhibits more power than the curve k=7 which has a pooled variance equal to 4. This demonstrates that the degree of violation of the assumption has little to do with determining the power of the test and that the pooled variance is the critical element in this determination.



GRAPH 2

COMPARISON OF POWER FOR VIOLATION OF ASSUMPTION vs. POOLED VARIANCE (pv)

EQUAL SAMPLE SIZES
$n_x = n_y = 10$

$k = \sigma_x^2/\sigma_y^2$
$K = 1, \quad \sigma_y^2=\sigma_x^2$

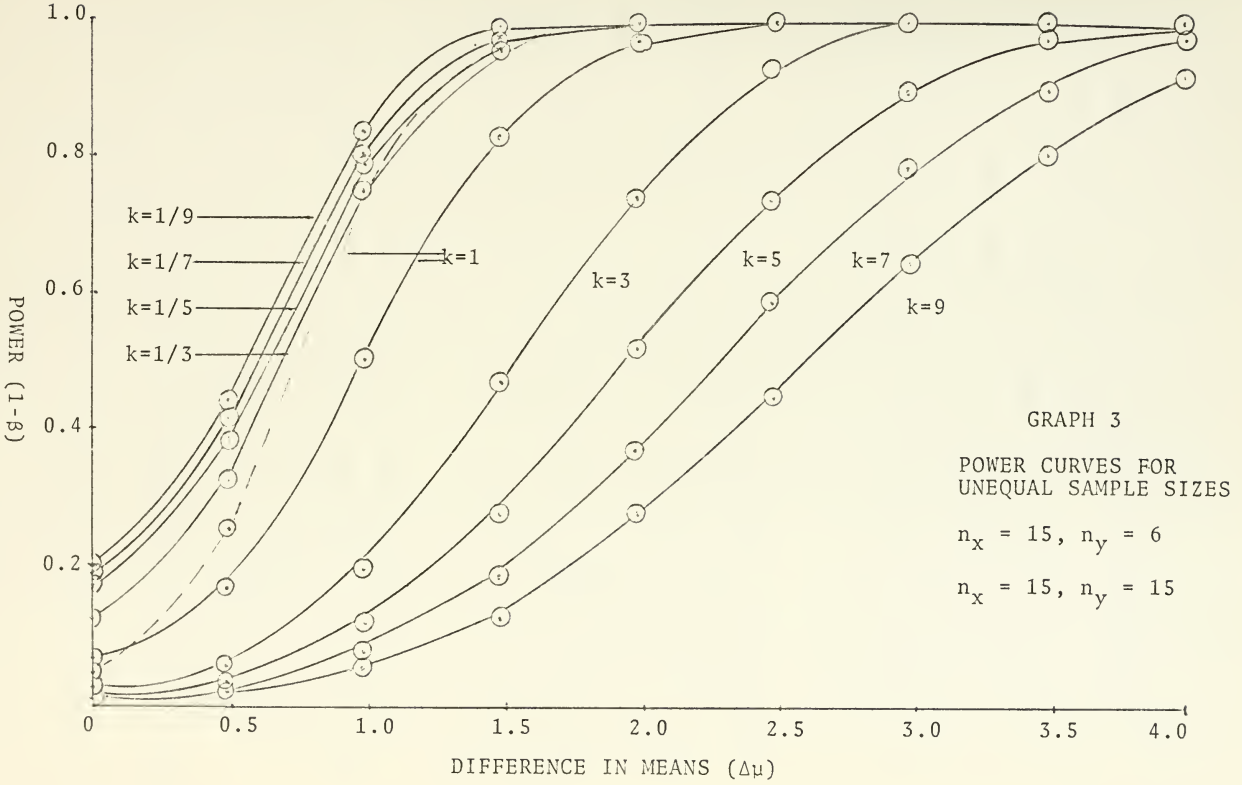
In all of the equal sample size cases the larger the sample size the greater the power of the test for an equal value of the pooled variance. This is a well documented result.

2. Unequal Sample Sizes

Welch (29) has written that a strong bias exists in the t-test, when the assumption of equal variance is violated, and the samples are not equal. This bias has been shown in the results of the estimated "true" rejection rates above. This same bias carries over to the power of the t-test under the same circumstances.

Graph 3 shows the power curve which results for various k values, of unequal samples size fifteen and six. The k values were achieved by maintaining the variance of Y equal to one and allowing the variance of X to range from 1/9 to 9. As in the case of equal sample sizes the power of the test is a function of the size of the pooled variance.

It should be noted that in the range of k from 1/9 to 1/3 the power of the test is extremely high but is achieved at the expense of an increase in the fraction of Type I errors when the two population means are equal. Here exists a good example of the conflict that develops when the fraction of Type II errors is decreased to the point where the rate of Type I errors becomes unacceptable. For k in the range 3 to 9 the power decreased with an increase in the fraction of Type II errors and as a



GRAPH 3

POWER CURVES FOR
UNEQUAL SAMPLE SIZES

$n_x = 15$, $n_y = 6$

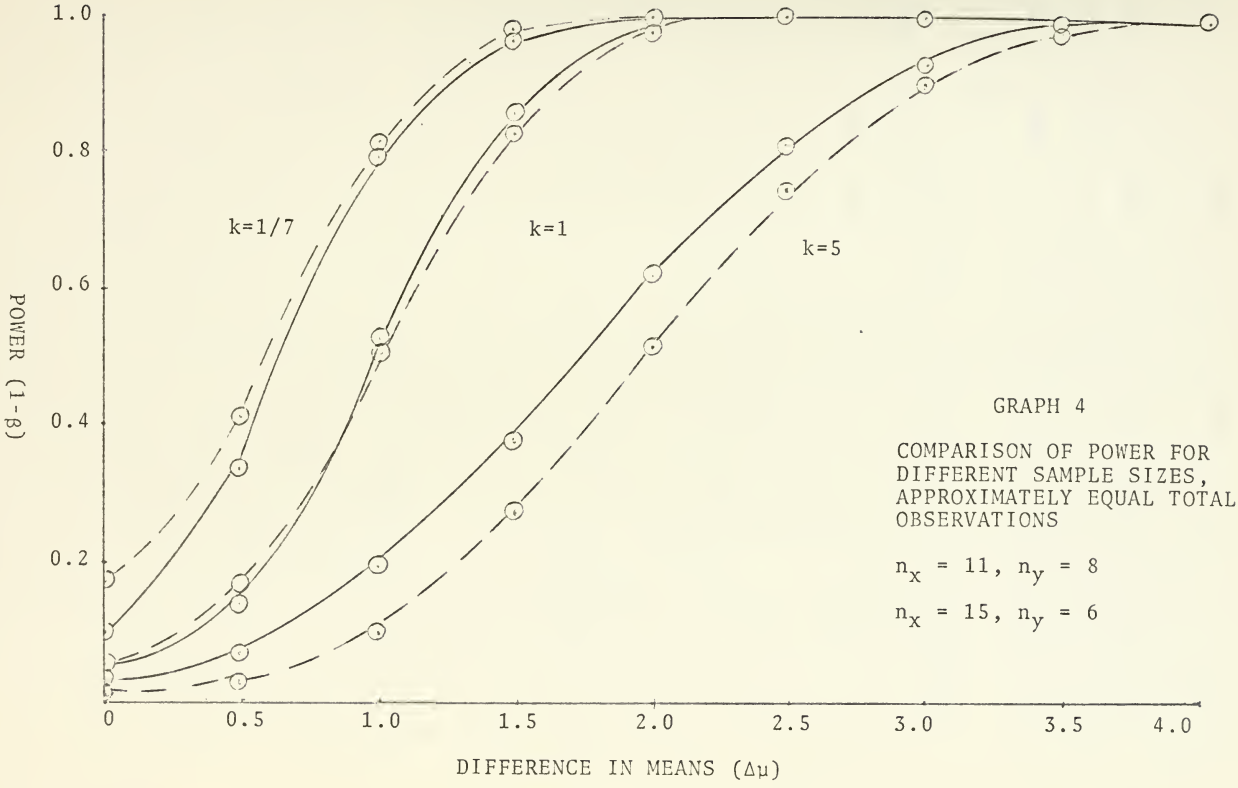
$n_x = 15$, $n_y = 15$

consequence the Type I errors decreased to a point where the rejection rate becomes significantly different from the expected rate. Similar results were obtained for the other unequal samples tested.

Also included in Graph 3 is a plot of the power curve for equal sample sizes n_x=15, n_y=15, and k=1. In comparing this curve to the similar k=1 curve for n_x=15, n_y=6, it can be seen that the power decreased because of the loss of information due to the fewer observations obtained for the Y population.

Graph 4 shows two cases where the total number of observations from both populations is about equal, but the difference between the sample sizes is not equal. In one case the total number of observations is 19 with n_x=11 and n_y=8, the difference between sample sizes being three. In the second case the total number of observations is 21 with n_x=15 and n_y= 6 and, therefore, the difference between sample sizes is 9.

For k=1 both cases have equal pooled variances and the power curves are almost identical. For k=1/7 the case n_x=15, n_y=6 has a smaller pooled variance than the case n_x=11, n_y=8 and as a result has a slightly higher power curve. For k=5 the relative size of the pooled variances is reversed and as a consequence the power curves are also reversed.



GRAPH 4

COMPARISON OF POWER FOR DIFFERENT SAMPLE SIZES, APPROXIMATELY EQUAL TOTAL OBSERVATIONS

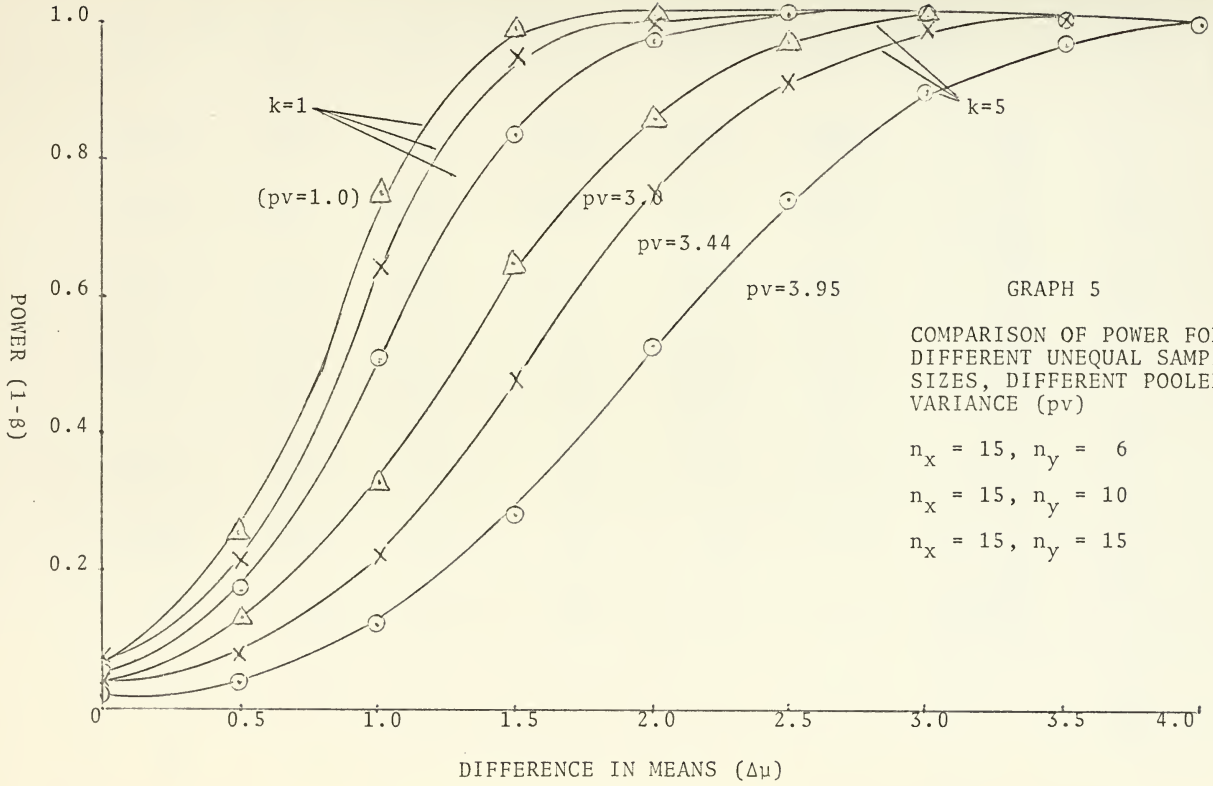
$n_x = 11$, $n_y = 8$

$n_x = 15$, $n_y = 6$

In Graph 5 three different cases are compared. For k=1 each of the cases has a pooled variance equal to one but the power curves are not identical because the total number of observations in each case are not equal. As the number of observations decreases, the power also decreases.

As the degree of violation of the assumption was increased to k=5 the pooled variance in each case is no longer equal. For n_x=15, n_y=6, the pooled variance is 3.95; n_x=15, n_y=10 the pooled variance is 3.44, and for n_x=15, n_y=15 the pooled variance is 3.03. At k=5 the relative relationship of the three power curves has changed somewhat from the case k=1. Under a changing degree of violation of the assumption a larger number of total observations causes a less rapid growth in the size of the pooled variance. This in turn results in a less rapid deterioration of the power of the test with an increasing degree of violation.

In all cases the power changed as a function of the size of the pooled variance. The same conclusion as was made in the case of equal sample sizes can be made here, that the power of the test is a function of the pooled variance rather than a function of the violation of the assumption of equality of variances. For unequal sample sizes though, the violation of the assumption causes



GRAPH 5

COMPARISON OF POWER FOR DIFFERENT UNEQUAL SAMPLE SIZES, DIFFERENT POOLED VARIANCE (pv)

$n_x = 15$, $n_y = 6$

$n_x = 15$, $n_y = 10$

$n_x = 15$, $n_y = 15$

a marked bias, and this is reflected in the power curves by either an increase or decrease in the α region of the curve at the point where the population means are equal.

Table 9
RESULTS FOR THE POWER OF THE t-TEST FOR
EQUAL SAMPLE SIZES AND VARIOUS k VALUES

n_x	n_y	$\Delta\mu$ 0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
k=1/9												
5	5	.069	.182	.477	.792	.953	.994	1.0				
10	10	.058	.276	.697	.929	.999	1.0					
15	15	.056	.429	.942	1.0							
k=1/7												
5	5	.066	.176	.458	.784	.950	.992	.999	1.0			
10	10	.057	.243	.694	.917	.999	1.0					
15	15	.055	.431	.941	.999	1.0						
k=1/5												
5	5	.056	.160	.453	.763	.944	.988	.999	1.0			
10	10	.054	.220	.675	.901	1.0						
15	15	.055	.408	.930	1.0							
k=1/3												
5	5	.056	.156	.398	.715	.919	.987	.998	1.0			
10	10	.054	.215	.655	.850	.998	1.0					
15	15	.053	.369	.896	.999	1.0						
k=1												
5	5	.049	.110	.291	.544	.784	.929	.985	.999	1.0		
10	10	.054	.190	.570	.889	.989	.996	1.0				
15	15	.049	.253	.752	.976	1.0						
k=3												
5	5	.054	.085	.177	.324	.514	.697	.836	.925	.970	.992	.998
10	10	.055	.127	.333	.616	.845	.960	.999	1.00			
15	15	.051	.169	.460	.794	.961	.996	1.0				

		Δμ										
n_x	n_y	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
k=5												
5	5	.060	.084	.145	.240	.370	.526	.675	.796	.885	.937	.975
10	10	.057	.100	.234	.448	.682	.859	.947	.990	1.0		
15	15	.054	.131	.330	.647	.858	.971	.994	1.0			
k=7												
5	5	.066	.077	.128	.207	.295	.425	.557	.675	.787	.872	.931
10	10	.056	.091	.187	.351	.567	.753	.889	.958	.982	.990	.999
15	15	.049	.114	.267	.517	.746	.907	.976	.995	1.0		
k=9												
5	5	.062	.080	.124	.183	.268	.360	.485	.598	.699	.789	.869
10	10	.059	.083	.174	.301	.471	.653	.805	.904	.965	.989	.996
15	15	.051	.098	.223	.436	.651	.841	.944	.985	.995	1.0	

Table 9 continued

Table 10
RESULTS FOR THE POWER OF THE t-TEST FOR
UNEQUAL SAMPLE SIZES AND VARIOUS k VALUES

n_x	n_y	Δμ 0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
k=1/9												
8	6	.094	.277	.671	.929	.994	.999	1.0				
10	6	.123	.337	.746	.957	.997	.999	1.0				
13	6	.165	.394	.805	.976	.999	1.00					
15	6	.190	.433	.831	.981	.999	1.00					
15	12	.084	.425	.925	.998	1.00						
11	8	.094	.339	.800	.982	.999	1.00					
k=1/5												
8	6	.087	.242	.636	.912	.993	.999	1.0				
10	6	.106	.301	.701	.941	.994	1.00					
13	6	.127	.361	.759	.968	.999	1.00					
15	6	.163	.385	.798	.974	.999	1.00					
15	12	.071	.407	.893	.998	1.00						
11	8	.083	.311	.761	.976	.999	1.00					
k=1/3												
8	6	.075	.218	.590	.887	.987	.999	1.0				
10	6	.084	.265	.656	.933	.993	.999	1.0				
13	6	.107	.305	.710	.955	.996	1.00					
15	6	.113	.318	.747	.963	.999	1.00					
15	12	.065	.356	.868	.995	1.00						
11	8	.071	.280	.725	.962	.999	1.00					
k=3												
8	6	.043	.075	.196	.423	.660	.834	.949	.985	.996	.999	1.0
10	6	.029	.075	.210	.436	.694	.872	.969	.993	.999	1.0	
13	6	.018	.060	.194	.439	.728	.903	.978	.997	1.00		
15	6	.016	.049	.187	.468	.736	.930	.985	.998	1.00		
15	12	.036	.133	.397	.749	.933	.992	.999	1.00			
11	8	.032	.094	.271	.539	.812	.942	.989	.999	1.00		
k=5												
8	6	.037	.060	.150	.280	.471	.666	.808	.916	.967	.992	.999
10	6	.024	.051	.126	.277	.490	.676	.849	.939	.982	.994	.999
13	6	.015	.037	.120	.279	.506	.727	.884	.962	.992	.998	1.0
15	6	.011	.030	.108	.276	.518	.743	.909	.971	.995	.999	1.0
15	12	.036	.090	.267	.555	.794	.948	.992	.999	1.00		
11	8	.036	.067	.193	.384	.623	.809	.925	.980	.996	.999	1.0

n_x	n_y	Δμ 0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
k=7												
8	6	.040	.058	.117	.220	.362	.526	.691	.819	.904	.962	.989
10	6	.024	.041	.095	.197	.363	.546	.724	.856	.936	.973	.990
13	6	.014	.026	.076	.190	.363	.576	.756	.884	.961	.988	.997
15	6	.008	.023	.072	.188	.374	.584	.778	.903	.970	.991	.999
15	12	.039	.080	.212	.418	.674	.854	.952	.991	.999	1.00	
11	8	.035	.053	.140	.296	.485	.696	.836	.932	.978	.995	.999
k=9												
8	6	.043	.053	.098	.181	.291	.436	.579	.736	.828	.902	.955
10	6	.022	.037	.079	.164	.288	.437	.605	.764	.865	.931	.970
13	6	.012	.024	.060	.146	.272	.454	.630	.798	.897	.960	.992
15	6	.007	.020	.053	.130	.277	.453	.653	.813	.920	.969	.993
15	12	.037	.075	.174	.352	.570	.765	.901	.965	.993	.999	1.00
11	8	.032	.056	.121	.248	.387	.580	.736	.872	.939	.979	.995

Table 10 continued

VI. SUMMARY AND CONCLUSIONS

This paper has investigated the robustness of the Student's t-test under violation of the assumption of the homogeneity of variances. The estimated "true" rejection rate and the estimated power of the test have been studied for the cases of equal and unequal sample sizes. Extensive use of computer simulation was made to conduct the study in each area of interest.

It was observed that the determination of the point at which the estimated "true" rejection rate became significantly different from the specified rate was dependent upon the criteria used. Two different criteria were established:

A. The total number of rejections at a single α level of 0.05.

B. The k value where the empirically generated distribution became significantly different from the tail of the t distribution.

It was also observed that the criteria became more stringent and difficult to satisfy from A to B. Consequently, for any case, the k critical intervals decreased when criteria B is applied instead of criteria A.

Table 11
LIMITS ON ROBUSTNESS OF t-TEST WITH RESPECT TO SCALE FACTOR VALUES, EQUAL SAMPLE SIZES

n_x	n_y	Criteria A	Criteria B
5	5	2.50-2.75	2.25-2.50
10	10	4.25-4.50	3.00-3.25
15	15	5.75-6.75	3.25-4.00

Concerning the estimated "true" rejection rates for "large" equal sample sizes of close to 15 observations each, it can be seen that even under the most stringent criteria, the ratio of the two population variances can be between 3.25 and 4.00 and the t-test will still provide an accurate statistical inference. Even at the small but equal sample sizes of five observations each, the magnitude of the variance ratio is great enough to imply that the t-test is fairly robust with respect to Type I rejection rates when the assumption of equality of variances is violated.

The test loses its robustness dramatically when sample sizes are unequal and a violation of equal variance occurs. Welch's predicitions have been verified by data generated by simulation. When the larger sample has the larger variance the difference between the two means tends to be underestimated and the estimated "true" rejection rate falls below the specified level. When the larger sample

has the smaller variance the difference between the two means tends to be overestimated and the estimated "true" rejection rate will be greater than the specified level.

With respect to power, the simulation has shown that the power of the test is a function of the pooled variance of the two populations and that it is not directly related to the degree of the violation of the assumption. This conclusion is valid for both equal and unequal sample sizes.

APPENDIX A

The following is a detailed description of the FORTRAN program used in this investigation. A sample program for a single case is contained on page 63.

The first term, IDUMMY = 0 is the beginning seed needed to activate the normal random number generator. The investigator must then enter the desired sample sizes for NX and NY. A quantity for the code name VAR is next read into the program. VAR is the standard deviation that has to be applied to one of the two samples to effect the desired variance ratio. The VAR value is printed on the computer output at this point in the program.

The value for the variable name DMEAN is next read into the computer. This value establishes the desired difference in population means used in studying the power of the test. In those instances when the estimated "true" rejection rate was investigated with the population means equal, DMEAN was set equal to 10. A DO loop is next entered and within each cycle of the DO loop, the variable names NUMACC, LPER10, LPER05, LPER02, LEER01, and LPER00, used to tabulate the empirical frequency distribution, were set equal to zero. In studying the power of the test, the DO loop incremented the difference in the population means by a factor of 0.5 for each cycle.

The program next enters the actual iteration DO loop which causes 50,000 different pairs of samples to be tested by the t-test. NX observations, drawn from a N(0,1) population, make up the sample representing the X population. Each of these observations is multiplied by the value VAR and then the value 10 is added. This causes the sample to appear to have been drawn from a N(10,VAR) population. NY observations are then drawn from the same initial N(0,1) population and make up the sample representing the Y population. To these observational values the value represented by the variable DMEAN, is added. This causes the NY sample to appear to have been drawn from a N(DMEAN,1) population.

With these two samples the t-test is then used to test the hypothesis that two population means are equal. The resulting absolute value of the observed t statistic is set equal to the variable name ATOBS. ATOBS is then compared against appropriate critical values of the t distribution. These appropriate critical values are functions of the desired α level, 0.10, 0.05, 0.02, 0.01, and 0.001 and the number of degrees of freedom for the samples being tested, NX + NY - 2. When the ATOBS value is greater than a particular critical value, a rejection of the null hypothesis occurs at that α level and the corresponding variable name associated with the particular α level of the empirical frequency distribution, is incremented by one.

The number of the 50,000 iterations in which the t-test concludes in accepting the null hypothesis, is tabulated by the variable name NUMACC. This is done for an α level of 0.05. From NUMACC the fraction of Type II errors is calculated and also the power of the test.

At the conclusion of 50,000 iterations for each case, NUMACC, β and the power of the test are printed out. Also the values of the empirical frequency distribution which have been developed from the test results are printed.

```
      DIMENSION X(15), Y(15)
      IDUMMY = 0
      NX = 15
      NY = 6
      DC 100 N = 1,9
      READ(5,5) VAR
. 5   FORMAT(F6.4)
      WRITE (6,21) VAR
21    FORMAT (//15X,F10.4//)
      DMEAN = 4.5
      DC 100 M = 1,11
      DMEAN = DMEAN + .5
      NUMACC = 0
      LPER10 = 0
      LPER05 = 0
      LPER02 = 0
      LPER01 = 0
      LPERCC = 0
      DC 50 I = 1,50(C0
      DC 10 J = 1, NX
10    X(J) = GRN(IDUMMY)*VAR + 10
      X(J) = X(J)
      DC 20 K = 1,NY
      Y(K) = GRN(IDUMMY)
20    Y(K) = Y(K) + DMEAN
      PCOLX = (NX-1)*(SX(X,NX,XBAR))**2
      PCCLY = (NY-1)*(SX(Y,NY,YBAR))**2
      TLOW1 = SQRT((PCOLX +POOLY)/(NX+NY-2))
      TLOW2 = SQRT((1.0/NX)+(1.0/NY))
      TOBS = (XBAR-YBAR)/(TLOW1*TLOW2)
      ATOBS = ABS(TOBS)
      IF (ATOBS .LT. 1.729) GO TO 30
      LPER10 = LPER10 + 1
      IF (ATOBS .LT. 2.093) GO TO 30
      LPER05 = LPER05 + 1
      IF (ATOBS .LT. 2.539) GO TO 50
      LPER02 = LPER02 + 1
      IF (ATOBS .LT. 2.861) GO TO 50
      LPER01 = LPER01 + 1
      IF (ATOBS .LT. 3.883) GO TO 50
      LPER00 = LPER00 + 1
      GO TO 50
  30  NUMACC = NUMACC + 1
  50  CONTINUE
      BETA = NUMACC/5000.0
      POWER = 1.0-BETA
      WRITE (6,601) NUMACC,BETA,POWER
60    FORMAT (110,10X,2F14.6)
      WRITE (6,61) LPER10,LPER05,LPER02,LPER01,LPER00
61    FORMAT (5110/)
100   CONTINUE
      STOP
      END
```

LIST OF REFERENCES

1. Bartlett, M.S., "The Information in Small Samples", Proceedings of the Cambridge Philosophical Society, v. 36, Part 4, p. 560-566, December 1936.

2. Chand, U., "Distribution Related to Comparison of Two Regression Coefficients", The Annuals of Mathematical Statistics, v. 21, no. 4, p. 507-521, December 1950.

3. Dixon, W. J., and Massey, F. J., Jr., Introduction to Statistical Analysis, 2d ed., McGraw-Hill, 1957.

4. Fisher, R. A., "Inverse Probability", Proceedings of the Cambridge Philosophical Society, v. 26, p. 528-535, October 1930.

5. Fisher, R. A., "The Fiducial Argument in Statistical Inference", Annals of Eugenics, v. 6, p. 391-398, 1935.

6. Fisher, R. A., "The Logic of Inductive Inference", Journal of the Royal Statistical Society, v. 98, p. 39-54, July 1935.

7. Fisher, R. A., "Uncertain Inference", Proceedings of the American Academy of Arts and Sciences, v. 71, p. 254-258, 1936.

8. Fisher, R. A., "A Note on Fiducial Inference", The Annals of Mathematical Statistics, v. 10, no. 4, p. 383-388, December 1939.

9. Gronow, D. G. C., "Test for the Significance of the Difference Between Means in Two Normal Populations Having Unequal Variances", Biometrika, v. 38, p. 252-256, June 1951.

10. Hopkins, J. W., and Clay, P. P. F., "Some Empirical Distributions of Bivariate T^2 and Homoscedasticity Criterion M Under Unequal Variance and Leptokurtosis", Journal of the American Statistical Association, v. 58, no. 304, p. 1048-1053, December 1963.

11. Keeping, E. S., Introduction to Statistical Inference, D. Van Nostrand Company, Princeton, New Jersey, 1962.

12. Kenny, J. F., and Keeping, E. S., Mathematics of Statistics, Part II, 2d ed., D. Van Nostrand Company, Princeton, New Jersey, Copy 1951.

13. Lehmann, E. L., Testing Statistical Hypothesis, John Wiley & Sons, Inc., New York, 1959.

14. Lindley, D. V., Introduction to Probability and Statistics, v. 2, Cambridge University Press, 1965.

15. Marsaglia, G., MacLaren, M. D., Bray, T. A., "A Fast Procedure for Generating Normal Random Variables", Communications of the ACM, v. 7, p. 4-10, January 1964.

16. McCullough, R. S., Gurland, J., and Rosenberg, L., "Small Sample Behavior of Certain Tests of the Hypothesis of Equal Means Under Variance Heterogeneity", Biometrika, v. 47, 3 and 4, p. 345-353, December 1960.

17. Meyer, P. L., Introductory Probability and Statistical Applications, Addison-Wesley Publishing Co., 1965.

18. Miller, I., and Freund, J. E., Probability and Statistics for Engineers, Prentice-Hall, Inc., 1965.

19. Murphy, B. P., "Some Two-Sample Tests When the Variances are Unequal: a Simulation Study", Biometrika, v. 54, 3 and 4, p. 679-683, December 1967.

20. Neyman, J., and Pearson, E. S., "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference", Biometrika, v. 20A, p. 175-240, 263-294, July 1928.

21. Neyman, J., and Pearson, E. S., "On the Problem of the Most Efficient Tests of Statistical Hypothesis", Philosophical Transactions of the Royal Society of London, v. 231 (series A), p. 289-337, March 1933.

22. Neyman, J., and Pearson, E. S., "Contributions to the Theory of Testing Statistical Hypothesis, Part I", Statistical Research Memoirs, v. 1, p. 1-37, 1936.

23. Owen, D. B., "The Power of the Student's t-Test", Journal of the American Statistical Association, v. 60, p. 320-333, March 1965.

24. Ray, W. D., and Pitman, A., "An Exact Distribution of the Fisher-Behrens-Welch Statistic for Testing the Difference Between the Means of Two Normal Populations With Unknown Variances", Journal of the Royal Statistical Society, v. 23, series B, p. 377-385, 1961.

25. Scheffe´, H., The Analysis of Variance, p. 331-369, John Wiley & Sons, Inc., New York, 1959.

26. Sverdrup, E., Basic Concepts of Statistical Inference, v. 1, North-Holland Publishing Co., Amsterdam, 1961.

27. Tuck, G. A., A Survey of Interval Estimation, M. S. Thesis, University of Oklahoma, Oklahoma City, 1964.

28. Weir, J. B., "Significance of the Difference Between Two Means When the Population Variances may be Unequal", Nature, v. 187, p. 438, July 1960.

29. Welch, B. L., "The Significance of the Difference Between Two Means When the Population Variances are Unequal", Biometrika, v. 29, p. 350-362, February 1938.

30. Welch, B. L., "The Generalization of 'Student's' Problem When Several Different Population Variances are Involved", Biometrika, V. 34, p. 28-35.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Asst Professor G. Tuck, Code Tk(thesis advisor) Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	1
4.	CAPT Harry A. Hadd, Jr., USMC Holmes Road Marine, Minnesota 55047	1
5.	Department of Operations Analysis, Code 55 Naval Postgraduate School Monterey, California 93940	1

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)
Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3 REPORT TITLE

Investigation of the Robustness of the Student's t-Test Under the Violation of the Assumption of Equality of Variances

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)
Master's Thesis; (December 1970)

5 AUTHOR(S) (First name, middle initial, last name)
Harry A. Hadd, Jr.

6 REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1970	69	30

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

The robustness of the Student's t-test is investigated under the violation of the assumption of equality of variances. With the aid of computer simulation, Type I and Type II error rates and the resulting statistical inference are studied and the effects of unequal variances on rejection rates and the power of the test are determined. Limits are determined on the degree of violation of the equality of variances that still leads to a satisfactory result when Student's distribution is used.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Student's t-Test						
Robustness						
Rejection Rates						

Thesis
H1068 Hadd
c.1 122451

Investigation of the robustness of the student's t-test under the violation of the assumption of equality of variances.